Wrangling Report

The wrangling process consists in 3 steps:

1. **Gather**
2. **Asses**
3. **Clean**

This is not a linear process, and several iterations were necessary to arrive at the final result.

1. Gather

The data sources used are:

- The WeRateDogs Twitter archive (twitter_archive_enhanced.csv) provided to students by Udacity.
  It contains basic tweet data (tweet id, timestamps, text, etc) for all tweets from WeRateDogs as they were on August 1st, 2017.

- The tweet image predictions, which include the breed of dog from the different tweets according to a neural network that uses the images to cast three predictions. This was also provided by Udacity to students, and it was downloaded programmatically.

- Twitter API to gather each tweet's retweet count and favorite count, among other things.

The data was gathered from these sources and saved as data frames in the Jupyter Notebook.

2. Assess

The data is assessed both visually and programatically using Jupyter Notebooks and Excel.

The quality and tidiness issues identified are:

Quality

A. Completeness

1. *Missing data in data frames*:
- In df: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.

- In twitter_counts_df: extended_entities, in_reply_to_status_id, in_reply_to_status_id_str, in_reply_to_user_id, in_reply_to_user_id_str, in_reply_to_screen_name, geo, coordinates, place, contributors, etc.

2. *Some tweets from df failed to match when querying the API for data*. Therefore, data in the twitter_counts_df for these tweet_ids will be missing.

B. Validity

3. *The id of the tweets appears in two columns as different data types in twitter_counts_df*.

4. *Some dog names are not valid dog names*, but rather words like "a", "just", "the", etc. These are either lower case text or text with less than 2 letters.

5. *Some of the tweets are retweets*.

6. *Some of the tweets are replies to tweets*.

7. *Some ratings have a denominator different to 10*. This was assessed by looking into the tweets on Twitter and investigating the reason for the strange numbers.

8. *Dog breed names appear in different ways* (e.g., 'miniature_pinscher', 'Shetland_sheepdog', Doberman) and some are not dog breeds.

C. Accuracy

9. *Some ratings numerators wrong or are suspiciously too high*. This was assessed by looking into the tweets on Twitter and investigating the reason for the strange numbers.

Tidiness

1. Each type of observational unit should form a single table, and we have 3 tables.

2. Dog stages should be one variable but instead its values are included in different columns.

3. tweet_id appears twice in the data from Twitter's API, as well as other variables from different tables, like 'in_reply_to_user_id'

3. Clean

After the assessment, the data was cleaned by defining the actions, coding and testing. The following actions were undertaken:

1. Remove tweets that are retweets or replies to tweets from the Twitter archive.
2. Remove tweets that failed to match in Twitter's API.
3. Remove tweets with no or irrelevant ratings.
4. Fill missing expanded urls.
5. Create a category for dog stages.
6. Rename and standardize invalid dog names.
7. Standardize dog ratings denominators and remove invalid ratings.
8. Create a single category for dog breeds.
9. Merge the data frames.
10. Remove empty, repeated and irrelevant columns, and fille any missing values

For an overview on the data analysis and visualization see the 'act_report.'
The whole process is included in the file 'wrangle_act.'